



## Estimation of Physical Properties of Amino Acids by Group-Contribution Method

Jhamb, Spardha Virendra; Liang, Xiaodong; Gani, Rafiqul; Hukkerikar, Amol Shivajirao

*Published in:*  
Chemical Engineering Science

*Link to article, DOI:*  
[10.1016/j.ces.2017.09.019](https://doi.org/10.1016/j.ces.2017.09.019)

*Publication date:*  
2018

*Document Version*  
Peer reviewed version

[Link back to DTU Orbit](#)

*Citation (APA):*  
Jhamb, S. V., Liang, X., Gani, R., & Hukkerikar, A. S. (2018). Estimation of Physical Properties of Amino Acids by Group-Contribution Method. *Chemical Engineering Science*, 175, 148-161.  
<https://doi.org/10.1016/j.ces.2017.09.019>

---

### General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

## Accepted Manuscript

Estimation of Physical Properties of Amino Acids by Group-Contribution Method

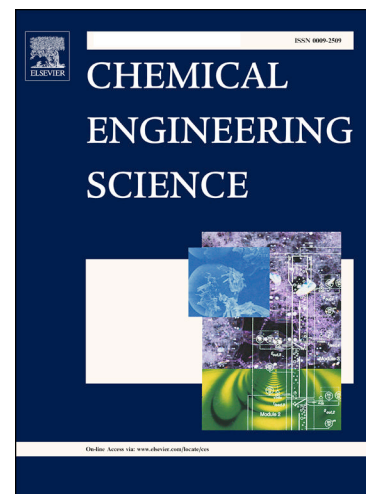
Spardha Jhamb, Xiaodong Liang, Rafiqul Gani, Amol Shivajirao Hukkerikar

PII: S0009-2509(17)30573-0  
DOI: <http://dx.doi.org/10.1016/j.ces.2017.09.019>  
Reference: CES 13796

To appear in: *Chemical Engineering Science*

Received Date: 24 May 2017  
Revised Date: 2 September 2017  
Accepted Date: 9 September 2017

Please cite this article as: S. Jhamb, X. Liang, R. Gani, A.S. Hukkerikar, Estimation of Physical Properties of Amino Acids by Group-Contribution Method, *Chemical Engineering Science* (2017), doi: <http://dx.doi.org/10.1016/j.ces.2017.09.019>



This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

# Estimation of Physical Properties of Amino Acids by Group-Contribution

## Method

Spardha Jhamb, Xiaodong Liang, Rafiqul Gani\*

Department of Chemical and Biochemical Engineering, Technical University of Denmark, DK-2800  
Lyngby, Denmark.

Amol Shivajirao Hukkerikar

Systems Engineering, Food Systems Business Unit, Alfa Laval Copenhagen A/S, Maskinvej 5, DK-  
2860, Søborg, Denmark.

## Abstract

In this paper, we present group-contribution (GC) based property models for estimation of physical properties of amino acids using their molecular structural information. The physical properties modelled in this work are normal melting point ( $T_m$ ), aqueous solubility ( $W_s$ ), and octanol/water partition coefficient ( $K_{ow}$ ) of amino acids. The developed GC-models are based on the published GC-method by Marrero and Gani (J. Marrero, R. Gani, Fluid Phase Equilib. 2001, 183-184, 183-208) with inclusion of new structural parameters (groups and molecular weight of compounds). The main objective of introducing these new structural parameters in the GC-model is to provide additional structural information for amino acids having large and complex structures and thereby improve predictions of physical properties of amino acids. The group-contribution values were calculated by regression analysis using a data-set of 239 values for  $T_m$ , 211 values for  $W_s$ , and 335 values for  $K_{ow}$ . Compared to other currently used GC-models, the developed models make significant improvements in accuracy with average absolute error of 10.8 K for  $T_m$  and logarithm-unit average absolute errors of 0.16 for  $K_{ow}$  and 0.19 for  $W_s$ .

**Keywords:** Normal melting point, Aqueous solubility, Octanol/water partition coefficient, Group contribution method, Amino acids.

## 1. Introduction

Amino acids, both natural as well as modified types, are widely used in the food and pharmaceutical industry on account of their nutritive and therapeutic properties. They are also extensively used in the cosmetics and personal care products owing to their high solubility and compatibility with the skin. Further, use of amino acids in consumer-oriented chemical products does not cause any harmful effects on the environment, a vital requirement from REACH® regulation (Echa.europa.eu, 2017) which compels European chemical companies to stop the use of hazardous chemical substances. Recently, driven by the demand for biodegradable chemicals, amino acids are also being studied for the development of biodegradable plastics. Table 1 lists selected amino acids and their applications in food, cosmetics and personal-care, therapeutic, and polymer industry.

The normal melting point ( $T_m$ ) of amino acids is a fundamental physical property of amino acids that allows estimation of solubility of amino acids (when used together with heat of fusion data as shown by Eq. (1) (Kontogeorgis and Folas, 2010)) and provides useful information on whether a substance is solid or liquid at normal conditions.

$$\ln x_i \gamma_i = -\frac{\Delta H_{fus}^\circ}{RT} \left(1 - \frac{T}{T_{m,i}}\right) + \frac{\Delta C_{p,i}}{R} \left(\frac{T_{m,i}-T}{T}\right) - \frac{\Delta C_{p,i}}{R} \ln \frac{T_{m,i}}{T} \quad (1)$$

where,  $x_i$  is the solubility of component  $i$  at temperature  $T$ ,  $\gamma_i$  is the activity coefficient,  $\Delta H_{fus}^\circ$  is the standard heat of fusion,  $R$  is the universal gas constant,  $T_{m,i}$  is the melting point and  $\Delta C_{p,i}$  is the difference between the heat capacity of the liquid and the solid.

The octanol/water partition coefficient ( $K_{ow}$ ) is the ratio of the equilibrium concentrations of organic compound in the two phases. By convention, the ratio of concentrations is expressed as octanol over water.  $K_{ow}$  is a measure of tendency of an organic compound to partition out

of water into other environmental compartments (these could be sediments, micro-organisms etc.). Organic compounds with  $K_{ow}$  less than  $10^2$  will not partition into any environmental compartments. Whereas, compounds with  $K_{ow}$  greater than  $10^6$  will tend to partition and accumulate in the environmental compartments.

Water solubility ( $W_s$ ), on the other hand, is defined as the maximum amount of an organic compound that is dissolved in pure water at a specific temperature, and it is also an important parameter in environmental studies. The relevance of  $T_m$ ,  $W_s$ , and  $K_{ow}$  of amino acids in chemical process-product design and environmental studies is therefore very high. There are three ways in which a property user can obtain these physical properties: (i) using available databases/open literature; (ii) performing laboratory measurements; and/or (iii) using property prediction methods. A key limitation associated with the use of databases is the limited number of compounds and limited amount of property-data stored in them. While use of experimentally measured physical properties is highly desirable, laboratory measurements may be time consuming, expensive, and sometimes may not even be feasible due to unavailability of samples or due to thermal decomposition issue. Therefore, it is more practical and convenient to employ property prediction methods to obtain the physical properties of amino acids, at least in the early stages of chemical product-process design (Constantinou and Gani, 1996).

The normal melting point of organic compounds is a difficult property to correlate because it is dependent upon the arrangement of the molecules in the crystal lattice as well as upon the strength of the pairwise group interactions (Katritzky et al., 2001).

Table 1. Selected applications of amino acids in food, cosmetics, personal-care, therapeutic, and polymer industry

Sr. No.	Amino acid	Amino acid derivative	Use	Reference
<b><i>Applications in food industry</i></b>				
1	Lysine	L-Lysine hydrochloride	Dietary supplement	(Leuchtenberger et al., 2005)
2	Glutamic acid	Monosodium Glutamate	Flavor enhancer	(Ault, 2004)
3	Tryptophan	L-Tryptophan	Antioxidant for preservation of milk powder	(Fox et al. 2006)
4	Cysteine	L-Cysteine	Antioxidant for preservation of fruit juices	(Lee, 1996)
5	Aspartic acid	Aspartyl-phenylalanine methyl ester	Low calorie artificial sweetener	(Tandel, 2011)
<b><i>Applications in cosmetics and personal-care products</i></b>				
1	Arginine	L-Arginine	Neutralizer in soap (alternative base)	(Sakamoto, 2006)
2	Glutamine	L-Glutamine	Neutralizer in soap (alternative acid)	(Sakamoto, 2006)
3	Proline	1-carbamimidoyl-L-proline	Improved hydration of stratum corneum (SC)	(Sakamoto, 2006)
4	Cystine	N-acetyl cysteine	Generation of melanin in UV-protecting creams	(Sakamoto, 2006)
5	Cysteine	L-Cysteine hydrochloride	Hair waving / Hair straightening agents	(Sakamoto, 2006)
6	Glycine	N-Acylglycinate	Mild, functional anionic surfactant	(Sakamoto, 2006)
7	Alanine	N-Acylalanate	Mild, functional anionic surfactant	(Sakamoto, 2006)
8	Serine	N-(4-pyridoxylmethylene)-L-serine	Anti-oxidative effects against UV Radiation	(Sakamoto, 2006)
<b><i>Therapeutic uses</i></b>				
1	Arginine	L-Arginine hydrochloride	Treat heart and circulatory diseases, combat fatigue	(Gad, 2010)
2	Phenylalanine	L-dihydroxyphenylalanine	Treatment of Parkinson's disease	(Hornykiewicz, 1974)
3	Tryptophan	5-hydroxytryptophan	Treatment of depression, anxiety and sleeping disorders	(Turner et al, 2006)
4	Glutamine	L-Glutamine	Treatment of mucositis, Gastrointestinal health	(Meletis et al., 2005)
5	L-Carnitine	N-Acetyl Carnitine	Treatment of Alzheimer's disease	(Meletis et al., 2005)
<b><i>As a solvent</i></b>				
1	Lysine	Potassium salt solution of L-Lysine	Solvent for CO <sub>2</sub> capture in flue gas	(Lerche et al. 2012)
<b><i>As a bio-degradable polymer (under investigation)</i></b>				
1	Alanine	DL-Alanine	Degradable polymeric drugs	(Domb, 1990)
2	Aspartic acid	Polyaspartate	Disposable diapers	(Low et al. 1996)

The modelling of melting point of amino acids is even more challenging and difficult due to the fact that these compounds are amphoteric and hence they exist as zwitterions at their isoelectric point (pI). Their ability to form zwitterions leads to stronger electrostatic attraction which in turn leads to larger energy requirement to break the bonds. For the estimation of normal melting point of organic compounds, several property prediction models have been published in the literature. These models can be broadly classified into two types: (i) GC based models; and (ii) Quantitative Structural-Property Relationship (QSPR) based models. In a GC-method, the property of a pure compound is a function of structurally dependent parameters, which are determined as a function of the frequency of the groups representing the pure compound and their contributions. In principle, it is an additive method, where the contributions of each group towards a property are summed to obtain the property value. Simamora and Yalkowsky (1994) proposed a simple GC method to estimate the normal melting point of aromatic compounds using 1690 compounds in the regression. Krzyzaniak et al. (1995) developed a simple GC method for estimation of normal melting of aliphatic, non-hydrogen bonding compounds. The GC methods developed by Joback and Reid (1987), Constantinou and Gani (1994), Marrero and Gani (2001), and Hukkerikar et al. (2012) also allow quick estimations of normal melting points of organic compounds without requiring substantial computational effort. The QSPR based models, on the other hand, describe a mathematical relationship between structural attributes (molecular descriptors) and a property of a set of pure compounds (Dearden, 2003). The QSPR models are based on specific classes of pure compounds (such as, aldehydes, alkanes, aromatics etc.). Katritzky et al. (1997) developed a correlation equation for the complete set of the melting points of 443 mono- and di-substituted benzenes. Godavarthy et al. (2006) developed a QSPR model for prediction of normal melting points of organic compounds based on the nonlinear descriptors and dataset of 1250 pure compounds. Zhao and Yalkowsky (1999) proposed a combined approach

involving group-contributions and molecular geometric parameters to estimate normal melting points of aliphatic compounds. A combination of the Joback and Reid's Method and Gold and Ogle Method has been used to develop the software program MPBPWIN in the EPI suite ®, which provides a recommended melting point based on the weighted average of the values predicted using the two methods (Jain and Yalkowsky, 2007). Table 2 provides a list of selected property prediction models for the estimation of normal melting point and their model performance statistics. As can be seen from Table 2, it is difficult to model normal melting points of organic compounds and achieve prediction accuracy close to the experimental measurement error.

For the estimation of  $W_s$ , Hukkerikar et al. (2012) developed a GC-model using a three-level parameter estimation approach (using a dataset of 4681 organic compounds) and this method requires only molecular structural information. There are several other GC-models available for estimation of aqueous solubility (Marrero and Gani, 2002; Klopman and Zhu, 2001; Kühne et al., 1995). For the estimation of  $K_{ow}$ , Marrero and Gani (2002) developed a GC-model using a dataset of 9560 organic compounds. Other well-known GC-methods for estimation of  $K_{ow}$  are those reported by Lin and Sandler (1999), Klopman and Wang (1994), and by Meylan and Howard (1996).

To the best of our knowledge, no property prediction models have been reported in the literature for the estimation of physical properties of amino acids (natural or modified types).

This work aims to develop property prediction models based on the Marrero and Gani GC-method (Marrero and Gani, 2001) to provide quick and reliable estimation of physical properties of amino acids. The paper first gives a brief overview of the Marrero and Gani GC-method; followed by model selection, development, and performance statistics; and finally, application of the developed GC-models. Tables containing list of groups (model



parameters) together with group-contributions (parameter values) are provided as a supporting information.

ACCEPTED MANUSCRIPT

Table 2. Selected property prediction models for estimation of normal melting point and their model performance statistics

Sr. No.	Method	Class of Pure Compounds	Data-points	R <sup>2</sup> <sup>a</sup>	SD <sup>b</sup> , K	AAE <sup>c</sup> , K	ARE <sup>d</sup>
<b>GC-models</b>							
1	Joback and Reid (1987)	Organic compounds	388	--	--	24.7	11.2
2	Constantinou and Gani (1994)	Organic compounds	312	--	18.28	14.03	7.23
3	Simamora and Yalkowsky (1994)	Aromatic compounds	1690	--	37.52	--	--
4	Krzyzaniak and Myrdal (1995)	Aliphatic, non-hydrogen-bonding compounds	596	0.977	34.3	--	--
5	Marrero and Gani (2001)	Organic compounds	1547	--	27.6	20.2	7.6
6	Qiang et al. (2009)	Organic compounds	730	--	--	14.46	6.67
7	Hukkerikar et al. (2012)	Organic compounds	5183	0.945	19.1	15.9	5.0
<b>QSPR models</b>							
1	Katritzky et al. (1997)	Mono- and di-substituted benzenes	443	0.8373	30.2	--	--
2	Godavarthy et al. (2006)	Organic compounds	1250	0.95	12.6	--	4.7
<b>Combined methods</b>							
1	Zhao and Yalkowsky (1999)	Aliphatic compounds	1040		34.4	--	20
2	Nigsch et al. (2006)	Organic compounds	4199	0.49	46.2	--	--
3	MPBPWIN in EPI ® suite	Organic compounds	2200	--	58.4	44.5	--
4	Admire and Yalkowsky (2015)	Polyhalogenated aliphatic and aromatic compounds	742	--	--	--	25.85

<sup>a</sup>R<sup>2</sup> = Coefficient of correlation; <sup>b</sup>SD = Standard deviation; <sup>c</sup>AAE = Average absolute error; <sup>d</sup>ARE = Average relative error

## 2. Methods and Tools Used for Property Modelling

To develop a GC-model that can provide accurate and reliable estimates of physical properties of amino acids, a systematic methodology based on the property-data-model-analysis is employed (Hukkerikar et al., 2013). It includes following steps:

- *Step-1:* Analysis of experimental values of property of various classes of amino acids as a function of their molecular weight to study the trend. This assists the model developer to select appropriate form of the property model function.
- *Step-2:* Select/verify the most appropriate form of the property model equation,  $f(x)$ , based on the observed trend.
- *Steps-3:* Perform parameter regression using all the available experimental data-points to obtain model parameter values (group contributions) which in turn can be used to predict the values of the physical property. Next, analysis of prediction errors (that is, the difference between the experimental value and the predicted value) for all amino acids in the data-set is made to identify amino acids and structural parameters involved in these amino acids having large prediction errors. The molecular structures of identified amino acids are analyzed and new structural parameters (functional groups) are included in the GC-model to describe well the structure of amino acids having large prediction errors. The purpose of introducing new structural parameters is to improve the model prediction performance through better correlation of experimental data. Regression analysis is performed again after including new structural parameters to obtain GC-model parameter values and model performance statistics such as, coefficient of determination ( $R^2$ ), standard deviation (SD), average absolute error (AAE), and average relative error (ARE). A description of these model performance statistics indicators is given in Section 2.5. If the calculated model performance statistics are acceptable then the estimated GC-model parameters are

considered as final. Otherwise, the procedure described is repeated until the desired correlation statistics are achieved.

While developing property models, very often, the experimental data-set is divided into training and validation sets. This should not be applicable for GC-models, since the validation set usually formed by randomly selecting the experimental data-points will exclude some of the GC-model parameters and thereby limit the application range of the GC-model. Also, when all the available experimental data-points are used in the parameter estimation step, it results in lower uncertainties of estimated model parameters and hence lower uncertainties (better reliability) of predicted property values (Hukkerikar et al., 2012).

## 2.1 Marrero and Gani (MG) Method

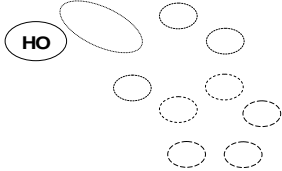
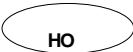
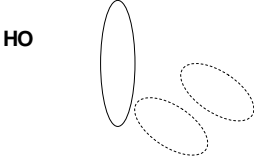
In the MG GC-method, the property estimation is performed at three levels. The first level has a large set of simple groups that allow for the representation of a wide variety of organic compounds. The second level of estimation involves groups that provide better description of proximity effects and can differentiate among isomers. The third level estimation includes groups that provide more structural information about molecular fragments of compounds whose description is insufficient through the first- and second-order groups; hence, this level allows estimation of complex heterocyclic and poly-functional acyclic molecules. The MG GC-model has the form (Marrero and Gani, 2001),

$$f(X) = \sum_i N_i C_i + w \sum_j M_j D_j + z \sum_k O_k E_k \quad (2)$$

The function  $f(X)$  is a function of property  $X$  and it may contain additional adjustable model parameters (universal constants) depending on the property involved. In Eq. (2),  $C_i$  is the contribution of the first-order group of type- $i$  that occurs  $N_i$  times.  $D_j$  is the contribution of the second-order group of type- $j$  that occurs  $M_j$  times.  $E_k$  is the contribution of the third-order group of type- $k$  that has  $O_k$  occurrences in a component. Eq. (2) is a general model for all the properties and the definition of  $f(X)$  is specific for each property  $X$ .

An example showing the representation of the molecular structure of the compound: 5-Hydroxymethylindane (CAS No. 51632-06-5) using first-order, second-order and third-order groups of the MG GC-method is given in Table 3.

Table 3. Representation of molecular structure of 5-Hydroxymethylindane using MG groups

First-order groups <sup>a</sup> / their occurrences	Second-order groups <sup>a</sup> / their occurrences	Third-order groups <sup>a</sup> / their occurrences
aCH /3	aC-CH <sub>n</sub> -OH (n in 1..2) /1	aC-CH <sub>ncyclic</sub> (fused rings) (n in 0..1)/2
aC /2	—	—
aC-CH <sub>2</sub> /1	—	AROM.FUSED[2]S <sup>2</sup> /1
OH /1	—	—
CH <sub>2</sub> (cyc) /3	—	—
		

<sup>a</sup>The marked circles in the drawing of the structure of 5-Hydroxymethylindane show the listed groups in respective symbols. The S<sup>2</sup> in the group AROM.FUSED[2]S<sup>2</sup> represents the position of branched fragment in the fused ring structure.

## 2.2 Data Sets

In the present study, we used a data-set containing 239 values for  $T_m$  of amino acids, 211 values for  $W_s$  of amino acids, and 335 values for  $K_{ow}$  of amino acids to build predictive GC-models. These experimentally measured property values of amino acids are collected from Sigma-Aldrich (Sigma-Aldrich, 2017) and from the database available at KT-Consortium. Table 4 provides description of data-sets consisting of natural as well as modified amino acids. The quantity of experimental data used in the development of property model influences the quality of parameter estimation and hence the prediction accuracy of that property model. Hukkerikar et al. (2012) discussed the effect of quantity of experimental data on the quality of parameter estimation and illustrated that by including all of the available experimental data of the property in the regression it is possible to improve the predictive

capability and application range of the property model. Therefore, in this work we have considered all of the available experimental data of properties of amino acids for regression purpose.

Table 4. Description of the data-sets used for the regression purpose

Derivatives of following amino acids	Data-points of $T_m$	Data-points of $W_s$	Data-points of $K_{ow}$
L-Alanine	13	8	5
$\beta$ -L-Alanine	4	3	1
L-Arginine	9	2	1
L-Asparagine	7	1	2
L-Aspartic acid	10	4	2
L-Cysteine	21	8	3
L-Glutamine	5	1	3
L-Glutamic acid	14	4	4
Glycine	27	25	35
L-Histidine	10	3	1
L-Isoleucine	5	4	2
L-Leucine	8	8	2
L-Lysine	10	2	3
L-Methionine	4	2	1
L-Ornithine	2	1	2
L-Proline	15	5	7
L-Phenylalanine	17	3	6
L-Serine	14	5	4
L-Threonine	8	1	1
L-Tyrosine	12	10	24
L-Tryptophan	8	2	2
L-Valine	9	4	3
aminobenzoic acid	--	29	125
aminonaphthalene sulfonic acid	--	13	2
aminobenzenesulfonic acid	--	17	1
others	7	46	93

Total	239	211	335
-------	-----	-----	-----

### 2.3 Selection of suitable property model function

The basis for selecting an appropriate property model is the study of behaviour of the property values of amino acids with their increasing molecular weight. This is illustrated for the case of  $T_m$  property in Figure 1.

Figure 1 shows that the plot of  $T_m$  of different classes of amino acids increase/decrease asymptotically with the increasing molecular weight suggesting that the most appropriate property function for  $T_m$  of amino acids should be of the exponential form. In this work, the property model function for prediction of  $T_m$  of amino acid is selected as,

$$\exp\left(\frac{T_m}{T_{m0} + \alpha \times MW}\right) = \sum_i N_i C_i + \sum_j M_j D_j + \sum_k O_k E_k \quad (3)$$

In Eq. (3),  $T_m$  is normal melting point and  $T_{m0}$  and  $\alpha$  are adjustable model parameter.  $MW$  is the molecular weight of amino acids. Eq. (3) is a modified version of the property model for  $T_m$  presented by Marrero and Gani (2001). The molecular weight of amino acids has been included as an additional parameter in Eq. (3) to achieve better prediction accuracy.

It can be seen from Figure 1 that, contrary to other amino acids, the nonpolar amino acids with aliphatic side chain, Fmoc-L amino acids with aliphatic side chain, and Bmoc-amino acids with aliphatic side chain show decrease in  $T_m$  with increase in molecular weight. This is possibly due to the reason that for these three classes of amino acids, there is an increase in the positive inductive effect on the zwitterionic group of the amino acid with the increase in the length of side chain (or the molecular weight). Due to the increase in the inductive effect, less energy is required during the phase change from solid to liquid state and hence the amino acid with a higher molecular weight would melt at a lower temperature compared to an amino

acid belonging to the same class with shorter chain length (or molecular weight). In the case of non-polar aromatic side chain compounds there is a negative inductive effect and resonance effect on the zwitterionic group of the amino acid while in the case of hydrochlorides of non-polar aliphatic side chain amino acids the intermolecular ionic interactions take precedence over the positive inductive effect.

The modelling of  $T_m$  of these amino acids thus becomes very challenging and requires special treatment such as identification and inclusion of new structural parameters in the GC-model to achieve reliable predictions of  $T_m$  of such molecules.

Similar analyses (trends analysis shown in Figure 1) have been performed (not shown in this paper) to obtain a suitable form of the property model for  $W_s$  and  $K_{ow}$  and these are given in Eq. (4) and Eq. (5).

$$\text{Log } W_s - A_{\text{Log}W_s} - B_{\text{Log}W_s} \text{ MW} = \sum_i N_i C_i + \sum_j M_j D_j + \sum_k O_k E_k \quad (4)$$

$$\text{Log } K_{ow} - A_{\text{Log}K_{ow}} = \sum_i N_i C_i + \sum_j M_j D_j + \sum_k O_k E_k \quad (5)$$

In Eq. (4) and Eq. (5),  $A_{\text{Log}W_s}$ ,  $B_{\text{Log}W_s}$ , and  $A_{\text{Log}K_{ow}}$  are adjustable model parameters.



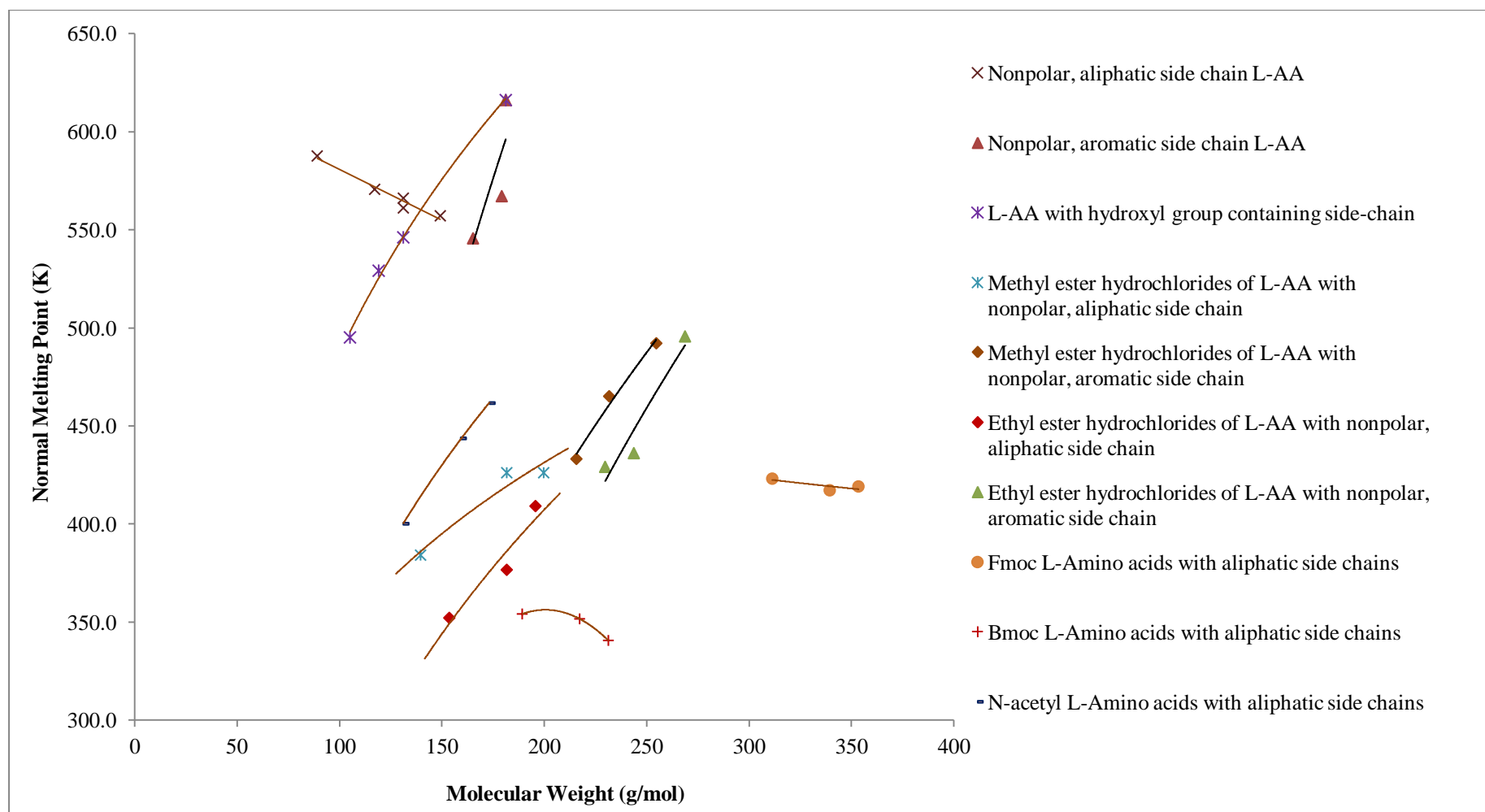


Figure 1.  $T_m$  of different classes of amino acids with their increasing molecular weight

ACCEPTED MANUSCRIPT

## 2.4 Parameter Estimation and Uncertainty Analysis

The optimization algorithm used for the parameter estimation is the Levenberg–Marquardt technique (Madsen et. al., 2004) and this algorithm was implemented in MatLab (The Mathworks, Natick, Massachusetts). The minimization of a cost function,  $S(\mathbf{P})$ , defined as the sum of the squares of the difference between the experimental value,  $X^{exp}$ , and evaluated property value,  $X^{pred}$ , provides the values of unknown parameters  $\mathbf{P}^*$ . This implies that  $\mathbf{P}^*$  is a set of model parameter values obtained at the minimum value of the cost function value.

$$S(\mathbf{P}) = \min \sum_{j=1}^N (X_j^{exp} - X_j^{pred})^2 \quad (6)$$

The subscript  $j$  indicates the pure compound evaluated and  $N$  is the total number of pure compounds included in the evaluation. After the estimation of the model parameters, uncertainty analysis is performed to quantify the model prediction errors. The methodology as discussed by Hukkerikar et al. (2012) is employed to estimate confidence interval of the predicted property value,  $X^{pred}$ , at  $\alpha_t$  significance level.

$$X_{1-\alpha_t}^{pred} = X^{pred} \pm \sqrt{\text{diag}(J(\mathbf{P}^*) \text{COV}(\mathbf{P}^*) J(\mathbf{P}^*)^T)} \cdot t(v, \alpha_t/2) \quad (7)$$

Where, the Jacobian matrix  $J(\mathbf{P}^*)$  calculated using  $\partial f / \partial \mathbf{P}^*$  represents the local sensitivity of the property model  $f$  to variations in the estimated parameter values  $\mathbf{P}^*$ . Eq. (7) is useful to assess the reliability of the prediction (when experimental data is available for the property).

If the experimental value is within the calculated confidence interval, then the property prediction method can be considered as reliable. When experimental data is unavailable, the calculated confidence interval provides a measure of the likely prediction error (uncertainty) of the predicted property value. This information can be used in the chemical product-process design to take into account the effects of uncertainties of predicted property values on the performance of the chemical product and/or process. For example, the product design of

liquid soap solution containing an amino acid as a surfactant requires a set of certain physical properties. If the experimental values of physical properties of that amino acid is not available, the chemist/product technologist can employ the developed GC based model to obtain the required physical properties together with the estimated prediction error (uncertainty). The prediction error can then be used to evaluate the effects of uncertainties in predicted property values on the performance (for example, lathering quality) of designed liquid soap solution.

## 2.5 Statistical Performance Indicators

The statistical significance of the developed correlations in this work is based on the following performance indicators (Hukkerikar et al., 2012).

- *Standard deviation* (SD): The root mean squared error value which is an estimator of standard deviation is calculated using,

$$SD = \sqrt{\sum_j (X_j^{exp} - X_j^{pred})^2 / N} \quad (8)$$

- *Average absolute error* (AAE): This is the measure of deviation of predicted property values from the experimentally measured property values and is given by,

$$AAE = \frac{1}{N} \sum_j |X_j^{exp} - X_j^{pred}| \quad (9)$$

- *Average relative error* (ARE): This provides an average of relative error calculated with respect to the experimentally measured property values and is given by,

$$ARE = \frac{1}{N} \sum_j |(X_j^{exp} - X_j^{pred}) / X_j^{exp}| \times 100 \quad (10)$$

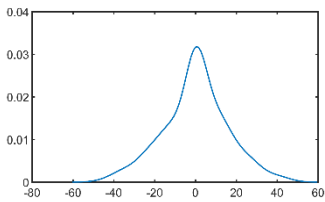
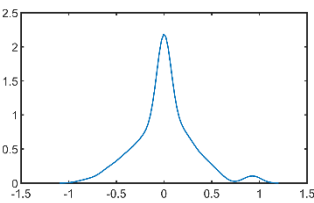
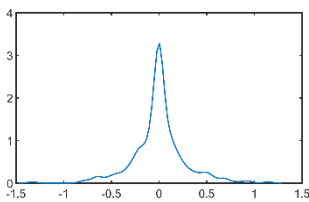
- *Coefficient of determination ( $R^2$ )*: This parameter provides information about the goodness of model fit. An  $R^2$  close to 1.0 indicates that the experimental data used in the regression have been fitted to a good accuracy. It is calculated using,

$$R^2 = 1 - \left[ \frac{\sum_j (X_j^{exp} - X_j^{pred})^2}{\sum_j (X_j^{exp} - \mu)^2} \right] \quad (11)$$

### 3. Model Performance Statistics

The model performance statistics for the developed GC-model for  $T_m$ ,  $W_s$ , and  $K_{ow}$  of amino acids are given in Table 5. In Table 5, N is the number of experimental data-points considered in the regression. The residuals ( $X^{exp} - X^{pred}$ ) for data-points considered in the regression are plotted in the form of residual distribution plot and is given in Table 5.

Table 5. Model performance statistics for developed GC-models

Property	N	$R^2$	Residual distribution plot	SD	AAE	ARE <sup>a</sup>	AE <sub>max</sub>
$T_m$	239	0.94		15.03	10.81	2.54	42.82
Log $W_s$	211	0.94		0.29	0.19	--	0.97
Log $K_{ow}$	335	0.99		0.25	0.16	--	1.18

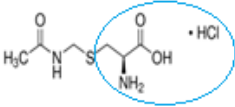
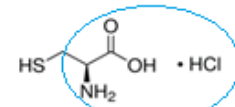
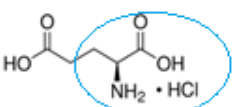
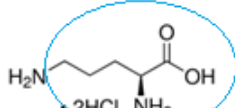
<sup>a</sup> ARE is not defined for Log $W_s$  and Log $K_{ow}$  since these properties have both positive and negative values.

In the final regression results for  $T_m$ , there are 123 amino acids out of 239 that have absolute error less than 10 K, 69 amino acids have absolute error ranging from 10 K to 20 K, and 47 amino acids have absolute error greater than 20 K. The maximum observed absolute error from regression is 42.8 K. Considering the difficulty in accurate prediction of normal melting points of amino acids, the obtained model performance statistics are acceptable. Further, the current GC-models are not capable of predicting the normal melting point of hydrochlorides of amino acids, thus limiting their application range. However, this developed GC-model can provide reliable estimates of normal melting point of both natural amino acids as well as modified types (salt forms). The new structural model parameters (that is, new third-order groups) are included in the developed GC-model to achieve better prediction accuracy. First, the Marrero and Gani GC-method (2001) is used as a reference to estimate the model parameter values and predicted  $T_m$  values to obtain absolute errors (deviation of predicted values from experimental values) for all amino acids. The molecular structures of amino acids having large deviations are now analysed and new structural model parameters are defined and included in the Marrero and Gani GC-model as additional new third-order groups, to provide more structural information to amino acids having large prediction errors. The definition of new third-order groups is based on the “molecular structure similarity criteria” approach. In this approach, the molecular structure of one amino acid is compared with the structures of other amino acids in the data-set to identify a set of amino acids that are “similar” in nature. Here, “similar” amino acids mean amino acids having one or more first-order groups (which are building blocks of molecule) in common. A list of new third-order groups together with their contributions is given in the supporting information (Table S3, groups 75 to 110). Table 6 indicates the improvement in predictions of  $T_m$  due to the new third-order group “ $(\text{CH}_n)_m\text{-CH}(\text{NH}_3+\text{Cl})\text{-COOH}$  ( $n$  in 0...2), ( $m$  in 0,1,2...)” for 4 amino

acids that are “similar” in molecular structure (the fragments of molecule that are “similar” in nature are marked as circles in Table 6).

ACCEPTED MANUSCRIPT

Table 6. An example illustrating improvement in the predictions of  $T_m$  using a new third-order group

Sl. No.	Amino acid (from data-set of $T_m$ )	Molecular structure with “similar” fragment marked in a circle	New structural parameter defined (as a new third-order group in GC-model)	Prediction error before introducing new structural parameter in GC-model for $T_m$			Prediction error after introducing new structural parameter in GC-model for $T_m$		
				Exp. value K	Prediction K	Error K	Exp. Value K	Prediction K	Error K
1	S-Acetamidomethyl-L-cysteine hydrochloride		“(CH <sub>n</sub> ) <sub>m</sub> -CH(NH <sub>3</sub> <sup>+</sup> Cl <sup>-</sup> )-COOH (n in 0...2), (m in 0,1,2...)	438.15	481.66	43.51	438.15	437.78	0.37
2	L-Cysteine hydrochloride			453.15	477.99	24.85	453.15	454.24	1.09
3	L-Glutamic acid hydrochloride			487.15	501.11	13.96	487.15	476.34	10.81
4	L-Ornithine dihydrochloride			471.15	515.75	44.60	471.15	486.84	15.69



For  $W_s$  and  $K_{ow}$  property, the goodness of the model-fits is very good and most of the data have been fitted to a good degree of accuracy with logarithm-unit average absolute errors is 0.17 for  $K_{ow}$  and 0.19 for  $W_s$ . In the final regression results for  $W_s$ , there are 192 amino acids (out of total 211) that have logarithm absolute error less than 0.5. For  $K_{ow}$ , there are 315 amino acids (out of total 335) that have logarithm absolute error less than 0.5. The classification of compounds on the basis of the absolute deviation from the experimental values of the three properties ( $T_m$ ,  $\text{Log } W_s$ ,  $\text{Log } K_{ow}$ ) have been shown in Figures 2, 3 and 4.

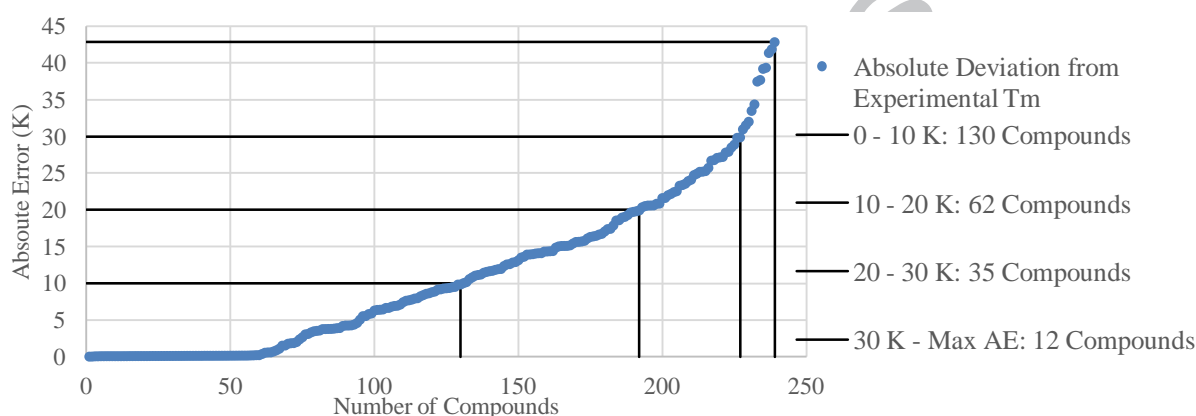


Figure 2. Classification of amino acids according to absolute deviation of  $T_m$

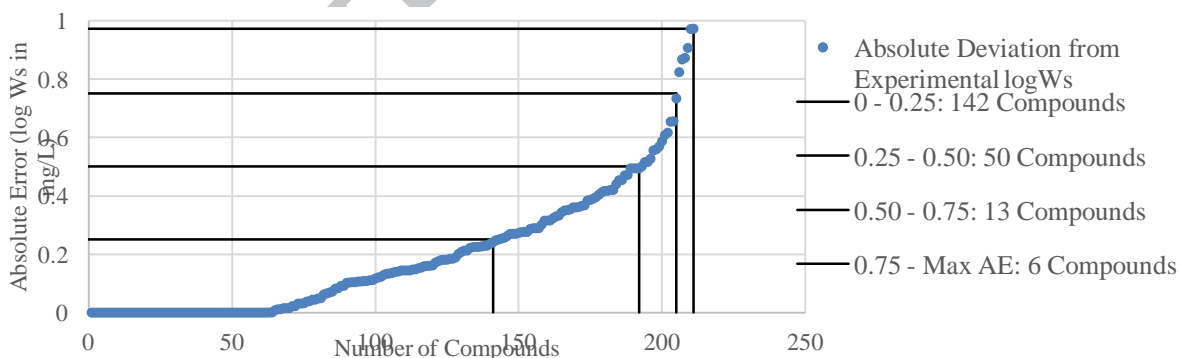


Figure 3. Classification of amino acids according to absolute deviation of  $\text{Log } W_s$

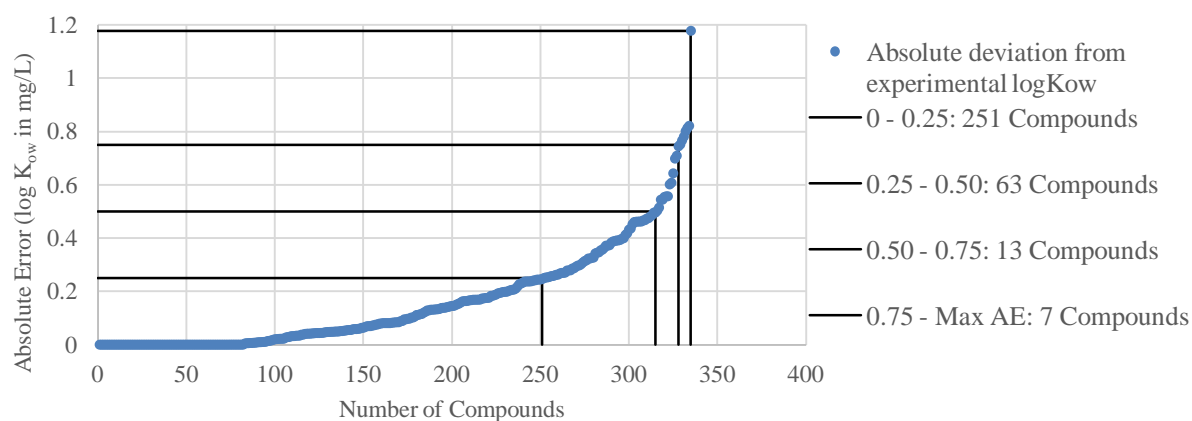


Figure 4. Classification of amino acids according to absolute deviation of  $\text{Log } K_{ow}$

The model performance statistics based on various types of amino acids is given in Table 7.

Table 7. Performance of developed GC-model for  $T_m$  for various types of amino acids

Derivatives of following amino acids	Data-points	SD K	AAE K	ARE %
L-Alanine	13	18.00	13.13	3.19
$\beta$ -Alanine	4	12.77	11.05	2.84
L-Arginine	9	16.18	11.64	2.61
L-Asparagine	7	15.98	13.73	2.98
L-Aspartic acid	10	13.65	11.63	2.69
L-Cysteine	21	13.06	9.41	1.99
L-Glutamine	5	24.36	21.52	5.07
L-Glutamic acid	14	11.93	9.39	2.32
Glycine	27	11.53	6.34	1.55
L-Histidine	10	9.26	5.79	1.16
L-Isoleucine	5	15.14	12.95	3.40
L-Leucine	8	19.23	10.44	3.47
L-Lysine	10	14.96	13.83	2.07
L-Methionine	4	28.67	17.57	6.58
L-Ornithine	2	12.34	27.44	2.43
L-Proline	15	15.08	11.83	2.89
L-Phenylalanine	17	13.58	10.22	2.46
L-Serine	14	11.26	8.76	2.05
L-Threonine	8	12.81	9.55	2.46

L-Tyrosine	12	15.56	11.40	2.57
L-Tryptophan	8	17.41	15.34	2.86
L-Valine	9	23.72	18.65	4.55
aminobenzoic acid	--	--	--	--
aminonaphthalene sulfonic acid	--	--	--	--
aminobenzene sulfonic acid	--	--	--	--
others	7	2.58	1.51	0.38
Total	239	--	--	--
Average	--	15.02	10.81	2.54

Table 8. Performance of developed GC-model for  $\log W_s$  for various types of amino acids

Derivatives of following amino acids	Data-points	SD $\log (mg/L)$	AAE $\log (mg/L)$
L-Alanine	8	0.17	0.12
$\beta$ -Alanine	3	0.20	0.16
L-Arginine	2	0.00	0.00
L-Asparagine	1	0.34	0.34
L-Aspartic acid	4	0.08	0.06
L-Cysteine	8	0.09	0.03
L-Glutamine	1	0.23	0.23
L-Glutamic acid	4	0.26	0.19
Glycine	25	0.18	0.13
L-Histidine	3	0.14	0.14
L-Isoleucine	4	0.17	0.17
L-Leucine	8	0.31	0.22
L-Lysine	2	0.22	0.16
L-Methionine	2	0.27	0.27
L-Ornithine	1	0.32	0.32
L-Proline	5	0.00	0.00
L-Phenylalanine	3	0.35	0.31
L-Serine	5	0.28	0.21

L-Threonine	1	0.17	0.17
L-Tyrosine	10	0.31	0.22
L-Tryptophan	2	0.18	0.13
L-Valine	4	0.20	0.12
aminobenzoic acid	29	0.35	0.27
aminonaphthalene sulfonic acid	13	0.35	0.22
aminobenzene sulfonic acid	17	0.50	0.42
others	46	0.25	0.15
Total	211	--	--
Average	--	0.29	0.19

Table 9. Performance of developed GC-model for  $\log K_{ow}$  for various types of amino acids

Derivatives of following amino acids	Data-points	SD $\log (mg/L)$	AAE $\log (mg/L)$
L-Alanine	5	0.13	0.07
$\beta$ -Alanine	1	0.24	0.24
L-Arginine	1	0.00	0.00
L-Asparagine	2	0.17	0.13
L-Aspartic acid	2	0.00	0.00
L-Cysteine	3	0.00	0.00
L-Glutamine	3	0.46	0.44
L-Glutamic acid	4	0.20	0.20
Glycine	35	0.29	0.22
L-Histidine	1	0.00	0.00
L-Isoleucine	2	0.05	0.03
L-Leucine	2	0.04	0.04
L-Lysine	3	0.21	0.12
L-Methionine	1	0.00	0.00
L-Ornithine	2	0.26	0.19
L-Proline	7	0.08	0.04
L-Phenylalanine	6	0.24	0.17
L-Serine	4	0.02	0.01

L-Threonine	1	0.03	0.03
L-Tyrosine	24	0.21	0.15
L-Tryptophan	2	0.00	0.00
L-Valine	3	0.27	0.27
aminobenzoic acid	125	0.29	0.20
aminonaphthalene sulfonic acid	2	0.10	0.10
aminobenzene sulfonic acid	1	0.00	0.00
others	93	0.22	0.12
Total	335	--	--
Average	--	0.25	0.16

#### 4. Application Examples

The application of developed GC-model to estimate  $T_m$  of amino acids is illustrated by considering the amino acid, L-Phenylalanine methyl ester hydrochloride (CAS: 7524-50-7). The experimentally measured value of  $T_m$  for L-Phenylalanine methyl ester hydrochloride is 433.2 K and this amino acid is present in the data-set used for parameter regression purpose.

Table 10. Prediction of  $T_m$  for L-Phenylalanine methyl ester hydrochloride (CAS: 7524-50-7)

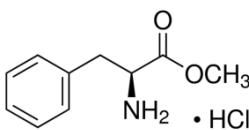
Compound:	Molecular structure	
L-Phenylalanine methyl ester		
hydrochloride		
Molecular formula:		
C6H5CH2CH(NH2)COOCH3 · HCl		
First-order groups	Occurrences	Group-contribution
CH <sub>3</sub>	1	0.6490
aCH	5	0.4154
aC-CH <sub>2</sub>	1	-0.1866
COO (except as above)	1	2.1259
-HCl	1	0.4434
Second-order groups	Occurrences	Group-contribution
--	--	--
Third-order groups	Occurrences	Group-contribution
CH(NH <sub>3</sub> +Cl <sup>-</sup> )-(CH <sub>n</sub> ) <sub>p</sub> -COO-	1	-0.8988
(CH <sub>2</sub> ) <sub>m</sub> (n in 1...2), (m,p in 0,1,2)		
$\exp\left(\frac{T_m}{T_{m0}+\alpha \cdot MW}\right)=\sum_i NiCi+\sum_j MjDj+\sum_l EkOk \quad \Rightarrow \quad T_m^{\text{pred}}=428.9 \text{ (Note: } T_{m0}=217.1367 \text{ K}$		
$\alpha=-0.0623 \text{ K mol g}^{-1}$ ). The experimental value of $T_m$ is 433.2 K. Hence, absolute deviation = 4.2 K		

Table 10 provides information of first-order, second-order, and third-order groups used to represent L-Phenylalanine methyl ester hydrochloride, their frequency (that is, occurrences in the structure) and the contributions for each group and taken from group-contribution tables given in the supporting information (Tables S1-S3). Using this information and the adjustable parameters, the predicted value of normal melting point is 430.9 (with absolute error =  $|433.2-428.9| = 4.2 \text{ K}$ ).

Table 10a. Covariance matrix  $COV(P^*)$  of model parameters comprising groups listed in Table 10 and adjustable model parameters

	$Tm0$	$A$	CH3	aCH	aC-CH2	COO (except as above)	HCl	CH(NH3+Cl-)-(CH2)n- COO-(CH2)m n in (0,...,2); m in (0,...,2)
$Tm0$	11723.97							
$A$	-5.35	0.004						
CH3	71.88	-0.01	0.89					
aCH	-53.78	0.02	-0.34	0.26				
aC-CH2	166.59	-0.07	1.23	-0.83	2.96			
COO (except as above)	-329.37	0.12	-2.65	1.53	-5.10	10.69		
HCl	-86.73	0.03	-0.70	0.39	-1.30	2.79	1.02	
CH(NH3+Cl-)-(CH2)n- COO-(CH2)m n in (0,...,2); m in (0,...,2)	167.42	-0.07	1.25	-0.79	2.63	-5.43	-1.48	3.06

Table 10b. Local sensitivity  $J(P^*)$  of  $T_m$  model with respect to model parameters comprising groups listed in Table 10 and adjustable model parameters

$\delta f / \delta Tm0$	$\delta f / \delta \alpha$	$\delta f / \delta CH3$	$\delta f / \delta aCH$	$\delta f / \delta aC-CH2$	$\delta f / \delta COO$ (except as above)	$\delta f / \delta HCl$	$\delta f / \delta CH(NH3+Cl-)-(CH2)n-COO-(CH2)m n in (0,...,2);m in (0,...,2)$
2.10	453.99	24.82	124.11	24.82	24.82	24.82	24.82

To calculate the confidence intervals, say the 95% confidence intervals of the predicted  $T_m$  value, the covariance matrix  $COV(P^*)$  given in Table 10a and the local sensitivity  $J(P^*)$  given in Table 10b are substituted in Eq. 7. For 95% confidence interval calculation, the  $t$ -distribution value corresponding to 0.05/2 percentile (i.e.  $\alpha_t/2$  percentile) and with 81 degrees

of freedom (obtained by deducting 158 estimated parameters from 239 data-points) is 1.96.

The calculated 95% confidence intervals of the estimated  $T_m$  value is,

$$T_{m(1-0.05)}^{pred} = T_m^{pred} \pm \sqrt{\text{diag}(J(\mathbf{P}^*)\text{COV}(\mathbf{P}^*)J(\mathbf{P}^*)^T)} \cdot t\left(v, \frac{\alpha_t}{2}\right)$$

$$T_{m(1-0.05)}^{pred} = 428.9 \pm 9.11 \times 1.96 = (430.9 \pm 17.85)K$$

It can be observed that the experimental value (433.2 K) falls in between the predicted confidence intervals indicating reliability of the developed model for  $T_m$ . Additional application examples to illustrate predictions of  $T_m$ ,  $\text{Log}W_s$  and  $\text{Log}K_{ow}$  of selected amino acids using developed GC-models are presented in Tables 11a-11c. The experimental values and the property estimation using MBPWIN & WSKOW of the EPI® suite are also given for comparison purpose.

Table 11a. Estimation of the  $T_m$  for selected amino acids

Compound-1	First-order groups	Occurrences ( $N_i$ )	Contribution ( $C_i$ )
N-(Carbobenzyloxy)- L-glutamic acid (CAS No. 1155-62-0)	CH2	2	-0.0687
	CH	1	-0.6259
	aCH	5	0.4154
	aC-CH2	1	-0.1866
	COOH	2	2.2226
	NHCO except as above	1	0.5721
	-O-	1	-0.8632
	Second-order groups	Occurrences ( $M_j$ )	Contribution ( $D_j$ )
	CHm(NHn)-COOH (m,n in 0..2)	1	3.3531
	aC-CHn-OOC (n in 1..2)	1	-0.6218
	Third-order groups	Occurrences ( $O_k$ )	Contribution ( $E_k$ )
	aC-CHn-OOC-NH-CHp-COOH (n,p in 1..2)	1	-1.0516
$\exp\left(\frac{T_m}{T_{m0} + \alpha MW}\right) = \sum_i N_i C_i + \sum_j M_j D_j + \sum_k O_k E_k \Rightarrow T_m = 387.3 \text{ K}$ The calculated 95% confidence interval of predicted property value = 25.0 K $T_{m,experimental} = 389.2 \text{ K}$			
Comparison of the predicted value with other methods:			
	Method	Predicted Value (K)	Absolute Error (K)



	Present work	387.3	1.9
	MPBPWIN	462.5	73.3

Compound-2	First-order groups	Occurrences ( $N_i$ )	Contribution ( $C_i$ )
L-Ornithine (CAS No. 70-26-8)	CH2	2	-0.0687
	COOH	1	2.2226
	CH2NH2	1	5.5543
	CHNH2	1	4.002
	Second-order groups	Occurrences ( $M_j$ )	Contribution ( $D_j$ )
	CH <sub>m</sub> (NH <sub>n</sub> )-COOH (m,n in 0..2)	1	3.3531
	Third-order groups	Occurrences ( $O_k$ )	Contribution ( $E_k$ )
	NH2-(CH2) <sub>m</sub> -CH(NH2)-COOH (m in 1,2...)	1	-6.2772
	$\exp\left(\frac{T_m}{T_{m0} + \alpha \times MW}\right) = \sum_i N_i C_i + \sum_j M_j D_j + \sum_k O_k E_k \Rightarrow T_m = 452.3 \text{ K}$ <p>The calculated 95% confidence interval of predicted property value = 36.0 K</p> <p><math>T_{m,experimental} = 413.2 \text{ K}</math></p>		
	Comparison of the predicted value with other methods:		
	Method	Predicted Value (K)	Absolute Error (K)
	Present work	452.3	39.1
	MPBPWIN	570.5	157.3

Compound-3	First-order groups	Occurrences ( $N_i$ )	Contribution ( $C_i$ )
L-Tyrosine tert-butyl ester (CAS No.16874-12-7)	CH3	3	0.649
	C	1	1.7561
	aCH	4	0.4154
	aC-CH2	1	-0.1866
	aC-OH	1	0.6147
	COO except as above	1	2.1259
	CHNH2	1	4.002
	Second-order groups	Occurrences ( $M_j$ )	Contribution ( $D_j$ )
	(CH3)3C	1	-1.8031
	AROMRINGS1s4	1	0.4342
	Third-order groups	Occurrences ( $O_k$ )	Contribution ( $E_k$ )
	CH(NH2)-(CH2) <sub>m</sub> -COO-(CH2) <sub>m</sub> (m in 1,2...)	1	-1.3093
	$\exp\left(\frac{T_m}{T_{m0} + \alpha \times MW}\right) = \sum_i N_i C_i + \sum_j M_j D_j + \sum_k O_k E_k \Rightarrow T_m = 450 \text{ K}$ <p>The calculated 95% confidence interval of predicted property value = 30.3 K</p> <p><math>T_{m,experimental} = 415.7 \text{ K}</math></p>		
	Comparison of the predicted value with other methods:		
	Method	Predicted Value (K)	Absolute Error (K)
	Present work	450	34.3
	MPBPWIN	390.7	25.0

Compound-4	First-order groups	Occurrences ( $N_i$ )	Contribution ( $C_i$ )
Methyl 6-aminohexanoate hydrochloride (CAS No. 1926-80-3)	CH3	1	0.649
	CH2	3	-0.0687
	CH2COO	1	0.3424
	CH2NH2	1	5.5543
	HCl (new)	1	0.4434
	$exp\left(\frac{T_m}{T_{m0} + \alpha \times MW}\right) = \sum_i N_i C_i + \sum_j M_j D_j + \sum_k O_k E_k \Rightarrow T_m = 394 \text{ K}$		
	The calculated 95% confidence interval of predicted property value = 31.1 K $T_{m,experimental} = 393.2 \text{ K}$		
Comparison of the predicted value with other methods:			
	Method	Predicted Value (K)	Absolute Error (K)
	Present work	394.0	0.8
	MPBPWIN	--	--

Table 11b. Estimation of the  $\text{Log}W_s$  for selected amino acids

Compound-1	First-order groups	Occurrences ( $N_i$ )	Contribution ( $C_i$ )
6-aminohexanoic acid (CAS No. 60-32-2)	CH2	4	-0.1437
	COOH	1	0.1727
	CH2NH2	1	1.1657
	Third-order groups	Occurrences ( $O_k$ )	Contribution ( $E_k$ )
	NHn-(CHn)m-COOH (m>2, n in 0..2)	1	2.1253
	$Log W_s = A_{LogWs} + B_{LogWs} MW + \sum_i N_i C_i + \sum_j M_j D_j + \sum_k O_k E_k = 5.4761$ The calculated 95% confidence interval of predicted property value = 0.8567 $LogW_{s,experimental} = 5.7033$		
Comparison of the predicted value with other methods:			
	Method	Predicted Value	Absolute Error
	Present work	5.4761	0.2271
	WSKOW	4.4082	1.2951
Compound-2	First-order groups	Occurrences ( $N_i$ )	Contribution ( $C_i$ )
N-aminocarbonyl $\beta$ -alanine (CAS No. 462-88-4)	CH2	2	-0.1437
	COOH	1	0.1727
	NH2CONH	1	1.8653
	Third-order groups	Occurrences ( $O_k$ )	Contribution ( $E_k$ )
	(aC)p-CO-NH-(CH2)m-COOH (m,p in 0,1,2,...)	1	0.2901
	$Log W_s = A_{LogWs} + B_{LogWs} MW + \sum_i N_i C_i + \sum_j M_j D_j + \sum_k O_k E_k = 4.6091$ The calculated 95% confidence interval of predicted property value		

$$LogW_{s.experimental} = 4.3201$$

Comparison of the predicted value with other methods:

Method	Predicted Value	Absolute Error
Present work	4.6091	0.2890
WSKOW	4.7929	0.4728

Compound-3	First-order groups	Occurrences ( $N_i$ )	Contribution ( $C_i$ )
2-amino 5-bromo benzenesulfonic acid (CAS No. 1576-59-6)	aCH	3	0.0270
	OH	1	1.2155
	aC-NH2	1	1.1577
	aC-Br	1	1.0651
	aC-SO2	1	1.1142
	Second-order groups	Occurrences ( $M_j$ )	Contribution ( $D_j$ )
	aC-SO2-OH	1	-0.6485
	AROMRINGs1s2s4	1	-0.3278
	Third-order groups	Occurrences ( $O_k$ )	Contribution ( $E_k$ )
	NH2-aC-aC-SO2-OH	1	-0.4725
$Log W_s = A_{Log W_s} + B_{Log W_s} MW + \sum_i N_i C_i + \sum_j M_j D_j + \sum_k O_k E_k = 3.3657$ The calculated 95% confidence interval of predicted property value = 1.1584 $Log W_{s, experimental} = 2.7745$			

Comparison of the predicted value with other methods:

	Method	Predicted Value	Absolute Error
	Present work	3.3657	0.5913
	WSKOW	3.3804	0.6059

Compound-4	First-order groups	Occurrences ( $N_i$ )	Contribution ( $C_i$ )
Methyl 4-methylamino benzenesulfonic acid (CAS No.24447-99-2)	CH3	1	0.2087
	aCH	4	0.0270
	OH	1	1.2155
	aC-NH	1	1.2708
	aC-SO2	1	1.1142
	Second-order groups	Occurrences ( $M_j$ )	Contribution ( $D_j$ )
	aC-SO2-OH	1	-0.6485
	AROMRINGs1s4	1	-0.1913
	$Log W_s = A_{LogWs} + B_{LogWs} MW + \sum_i N_i C_i + \sum_j M_j D_j + \sum_k O_k E_k = 4.5493$ The calculated 95% confidence interval of predicted property value = 1.1885 $LogW_{s,experimental} = 5.4533$		

Comparison of the predicted value with other methods:

Method	Predicted Value	Absolute Error
--------	-----------------	----------------

	Present work	4.5493	0.9040
	WSKOW	4.0162	1.4371

Table 11c. Estimation of the  $\text{Log}K_{ow}$  for selected amino acids

Compound-1	First-order groups	Occurrences ( $N_i$ )	Contribution ( $C_i$ )
2-(diethylamino)-2-oxoethyl ester benzoic acid (CAS No.64649-63-4)	CH3	2	0.0690
	CH2	1	0.4268
	aCH	5	0.1131
	aC-COO	1	0.8095
	CON(CH2)2	1	-0.6544
	$Log K_{ow} = A_{LogKow} + \sum_i N_i C_i + \sum_j M_j D_j + \sum_k O_k E_k = 1.98$		
	The calculated 95% confidence interval of predicted property value = 0.47		
$LogK_{ow,experimental} = 2.06$			
Comparison of the predicted value with other methods:			
	Method	Predicted Value	Absolute Error
	Present work	1.98	0.08
	WSKOW	1.98	0.08
Compound-2	First-order groups	Occurrences ( $N_i$ )	Contribution ( $C_i$ )
4-aminosulfonylbenzoic acid (CAS No. 138-41-0)	aCH	4	0.1131
	aC-COOH	1	0.3486
	NH2 except as above	1	-2.6336
	aC-SO2	1	-0.0373
	Second-order groups	Occurrences ( $M_j$ )	Contribution ( $D_j$ )
	AROMRINGS1s4	1	-0.1894
	aC-SO2NHn (n>=0;n<3)	1	1.402
$Log K_{ow} = A_{LogKow} + \sum_i N_i C_i + \sum_j M_j D_j + \sum_k O_k E_k = 0.04$			
The calculated 95% confidence interval of predicted property value = 0.48			
$LogK_{ow,experimental} = 0.5$			
Comparison of the predicted value with other methods:			
	Method	Predicted Value	Absolute Error
	Present work	0.04	0.46
	WSKOW	0.25	0.25
Compound-3	First-order groups	Occurrences ( $N_i$ )	Contribution ( $C_i$ )
p-bromophenyl glutamic acid (CAS No. N.A.)	CH2	2	0.4268
	CH	1	0.5554
	aCH	4	0.1131
	COOH	2	-1.9702
	aC-NH	1	0.4606
	aC-Br	1	1.1759
	Second-order groups	Occurrences ( $M_j$ )	Contribution ( $D_j$ )

	CHm(NHn)-COOH (m,n in 1 0..2)	1	0.8151
	AROMRINGS1s4	1	-0.1894
	$Log K_{ow} = A_{LogKow} + \sum_i N_i C_i + \sum_j M_j D_j + \sum_k O_k E_k=0.88$		
	The calculated 95% confidence interval of predicted property value = 0.62 $LogK_{ow,experimental} = 1.12$		
Comparison of the predicted value with other methods:			
	Method	Predicted Value	Absolute Error
	Present work	0.88	0.24
	WSKOW	1.34	0.22
Compound-4	First-order groups	Occurrences ( $N_i$ )	Contribution ( $C_i$ )
4-aminomethyl benzoic acid (CAS No. 56-91-7)	aCH	4	0.1131
	aC-CH2	1	1.0941
	aC-COOH	1	0.3486
	NH2 except as above	1	-2.6336
	Second-order groups	Occurrences ( $M_j$ )	Contribution ( $D_j$ )
	aC-CHn-NHm (n in 1..2; m in 0..2))	1	0.5538
	AROMRINGS1s4	1	-0.1894
	$Log K_{ow} = A_{LogKow} + \sum_i N_i C_i + \sum_j M_j D_j + \sum_k O_k E_k= 0.32$		
	The calculated 95% confidence interval of predicted property value = 0.71 $LogK_{ow,experimental} = 1.03$		
Comparison of the predicted value with other methods:			
	Method	Predicted Value	Absolute Error
	Present work	0.32	0.71
	WSKOW	0.95	0.08

## 5. Conclusions

Group-contribution based property model have been developed with the objective of providing reliable and quick estimation of physical properties ( $T_m$ ,  $W_s$  and  $K_{ow}$ ) of natural as well as modified amino acids for applications in the chemical process-product design. The developed GC-model requires only the molecular structural information of amino acids as an input. For both natural as well as modified unnatural amino acids, the developed GC-models performs very well and the performance statistics are better in comparison to currently used property prediction models. The use of the developed GC-models to estimate  $T_m$ ,  $W_s$  and  $K_{ow}$  of amino acids is illustrated through several application examples. The application range of

the developed GC-models is very broad: these can be used by process engineers to perform chemical product-process design and/or to perform environmental impact analysis. Motivated by the results obtained in this work, our current and future work is focused on the property modelling of heat of fusion of amino acids which will allow chemists/process technologists to estimate solubility of amino acids.

## Supporting Information

Following tables are included as a supporting information.

Table S1. List of first-order groups and their contributions  $C_i$  for the GC-model for  $T_m$ ,  $\log W_s$  and  $\log K_{ow}$  of amino acids

Table S2. List of second-order groups and their contributions  $D_j$  for the GC-model for  $T_m$ ,  $\log W_s$  and  $\log K_{ow}$  of amino acids

Table S3. List of third-order groups and their contributions  $E_k$  for the GC-model for  $T_m$ ,  $\log W_s$  and  $\log K_{ow}$  of amino acids

## Acknowledgments

This research was carried out as a part of the PhD project funded by KT Consortium, Department of Chemical and Biochemical Engineering, Technical University of Denmark (DTU) and in collaboration with industrial research partner Alfa Laval Copenhagen A/S, Denmark.

## References

1. Admire, B., Lian, B., Yalkowsky, S., 2015. Estimating the physicochemical properties of polyhalogenated aromatic and aliphatic compounds using UPPER: part 1. boiling point and melting point. *Chemosphere*. 119, 1436-1440
2. Ault, A., 2004. The Monosodium Glutamate Story: The commercial production of MSG and other amino acids, *Journal of Chemical Education*. 81(3), 347-355
3. Constantinou, L., Gani, R., 1994. New group contribution method for estimating properties of pure compounds. *AIChE Journal*. 40(10), 1697-1710
4. Dearden, J., 2003. Quantitative structure-property relationships for prediction of boiling point, vapor pressure and melting point. *Environmental Toxicology and Chemistry*. 22(8), 1696-1709
5. Domb, A.J., 1990. Biodegradable polymers derived from amino acids. *Biomaterials*. 11(9), 686-689
6. EPA.gov. (2017). EPI Suite™-Estimation Program Interface | Predictive models and tools for assessing chemicals under the toxic substances control act (TSCA) | US EPA. [online] Available at: <https://www.epa.gov/tsca-screening-tools/epi-suite-estimation-program-interface> [Accessed 18 Apr. 2017]
7. Echa.europa.eu. (2017). REACH - ECHA. [online] Available at: <https://echa.europa.eu/regulations/reach> [Accessed 18 Apr. 2017].
8. O'Connor, T.P., O'Brien, N.M., 2006. Lipid oxidation, in: Fox, P. F and McSweeney, P. L. H., (Eds.), *Advanced Dairy Chemistry*. third ed. Springer Science+ Business Media Inc., New York, 557-585
9. Gad, M.Z., 2010. Anti-aging effects of L-arginine. *Journal of Advanced Research*. 1(3), 169-177

10. Gani, R., Constantinou, L., 1996. Molecular structure based estimation of properties for process design. *Fluid Phase Equilibria*. 116(1-2), 75-86
11. Godavorthy, S., Robinson, R., Gasem, K., 2006. An improved structure–property model for predicting melting-point temperatures. *Industrial & Engineering Chemistry Research*. 45(14), 5117-5126
12. Hornykiewicz, O., 1974. The mechanisms of action of L-dopa in parkinson's disease. *Life Sciences*. 15.7, 1249-1259
13. Hukkerikar, A.S., Kalakul, S., Sarup, B., Young, D.Y., Sin, G., Gani, R., 2012. Estimation of environment-related properties of chemicals for design of sustainable processes: development of group-contribution+ (GC+) property models and uncertainty analysis. *Journal of Chemical Information and Modeling*. 52(11), 2823-2839
14. Hukkerikar, A.S., Meier, R.J., Gani, R., 2013. A method to estimate the enthalpy of formation of organic compounds with chemical accuracy. *Fluid Phase Equilibria*. 348, 23-32
15. Jain, A., Yalkowsky, S., 2007. Comparison of two methods for estimation of melting points of organic compounds. *Industrial & Engineering Chemistry Research*. 46(8), 2589-2592
16. Joback, K.G., Reid, R.C., 1987. Estimation of pure component properties from group-contributions. *Chemical Engineering Communications*. 57, 233-243
17. Lee, B. H., 1995. *Fundamentals of Food Biotechnology*, first ed. VCH Publishers Print.
18. Leuchtenberger, W., Klaus H., and Karlheinz D., 2005. Biotechnological production of amino acids and derivatives: current status and prospects. *Applied Microbiology and Biotechnology*. 69(1), 1-8



19. Low, K.C., Wheeler, A.P.; Koskan, L.P., 1996. Commercial poly(aspartic acid) and its uses, *Advances in Chemistry Series*. 248, 99-111
20. Katritzky, A., Maran, U., Karelson, M., Lobanov, V., 1997. Prediction of melting points for the substituted benzenes: a QSPR approach. *Journal of Chemical Information and Computer Sciences*. 37(5), 913-919
21. Katritzky, A., Jain, R., Lomaka, A., Petrukhin, R., Maran, U., Karelson, M., 2001. Perspective on the relationship between melting points and chemical structure. *Crystal Growth & Design*. 1(4), 261-265
22. Klopman, G., Li, J., Wang, S., Dimayuga, M., 1994. Computer automated log P calculations based on an extended group contribution approach. *Journal of Chemical Information and Modeling*. 34(4), 752-781
23. Klopman, G., Zhu, H., 2001. Estimation of the aqueous solubility of organic molecules by the group contribution approach. *Journal of Chemical Information and Computer Sciences*. 41(2), 439-445
24. Kontogeorgis, G. and Folas, G. 2010. *Thermodynamic models for industrial applications*. Hoboken, N.J: Wiley.
25. Krzyzaniak, J., Myrdal, P., Simamora, P., Yalkowsky, S., 1995. Boiling point and melting point prediction for aliphatic, non-hydrogen-bonding compounds. *Industrial & Engineering Chemistry Research*. 34(7), 2530-2535
26. Kühne, R., Ebert, R., Kleint, F., Schmidt, G., Schüürmann, G., 1995. Group contribution methods to estimate water solubility of organic chemicals. *Chemosphere*, 30(11), 2061-2077
27. Lerche, B.M., Stenby, E.H., Thomsen, K., 2012. CO<sub>2</sub> capture from flue gas using amino acid salt solutions (Doctoral dissertation, Technical University of Denmark, Department of Chemical and Biochemical Engineering)

28. Lin, S., Sandler, S., 1999. Prediction of octanol–water partition coefficients using a group contribution solvation model. *Industrial & Engineering Chemistry Research*. 38(10), 4081-4091
29. Madsen, K., Nielsen, H.B., Tingleff, O., 2004. *Methods for non-linear least squares problems*. second ed., Informatics and Mathematical Modelling, Technical University of Denmark, Lyngby
30. Marrero, J., Gani, R., 2001. Group-contribution based estimation of pure component properties. *Fluid Phase Equilibria*. 183-184, 183-208
31. Marrero, J., Gani, R., 2002. Group-contribution-based estimation of octanol/water partition coefficient and aqueous solubility. *Industrial & Engineering Chemistry Research*. 41(25), 6623-6633
32. Meletis, C.D., Barker, J.E., 2005. Therapeutic uses of amino acids. *Alternative and Complementary Therapies*. 11(1), 24-28
33. Meylan, W., Howard, P., Boethling, R., 1996. Improved method for estimating water solubility from octanol/water partition coefficient. *Environmental Toxicology and Chemistry*. 15(2), 100-106
34. Nigsch, F., Bender, A., Buuren, B.V., Tissen, J., Nigsch, E. and Mitchell, J., 2006. Melting point prediction employing k-nearest neighbor algorithms and genetic parameter optimization. *Journal of Chemical Information and Modeling*. 46(6), 2412-2422
35. Sakamoto, K., 2016. Amino acids and derivatives, in: Sivamani, R.K., Jagdeo J.R., Elsner, P., Maibach, H.I. (Eds.), *Cosmeceuticals and Active Cosmetics*, third ed. CRC Press., 163 - 175.

36. Sigma-Aldrich. (2017). biochemical-products. [online] Available at: <http://www.sigmaaldrich.com/life-science/biochemicals/biochemical-products.html?TablePage=16282308> [Accessed 18 Apr. 2017].
37. Simamora, P., Yalkowsky, S.H., 1994. Group contribution methods for predicting the melting points and boiling points of aromatic compounds. *Industrial & Engineering Chemistry Research*. 33(5), 1405-1409
38. Tandel, K.R., 2011. Sugar Substitutes: Health controversy over perceived benefits, *Journal of Pharmacology and Pharmacotherapeutics*. 2(4), 236-243
39. Turner, E.H., Jennifer M.L., Blackwell, A.D., 2006. Serotonin a la carte: supplementation with the serotonin precursor 5-hydroxytryptophan. *Pharmacology & Therapeutics*. 109(3), 325-338
40. Wang, Q., Ma, P., Neng, S., 2009. Position group-contribution method for estimation of melting point of organic compounds. *Chinese Journal of Chemical Engineering*. 17(3), 468-472
41. Zhao, L., Yalkowsky, S., 1999. A combined group contribution and molecular geometry approach for predicting melting points of aliphatic compounds. *Industrial & Engineering Chemistry Research*. 38(9), 3581-3584

## Abbreviations

AAE	average absolute error
ARE	average relative error [%]
$A_{LogWs}$	adjustable parameter of Eq. (4)
$A_{Logkw}$	adjustable parameter of Eq. (5)
$\alpha$	adjustable parameter of Eq. (3)
$B_{LogWs}$	adjustable parameter of Eq. (4)
$C_i$	contribution of first-order group of type- $i$
$COV(\mathbf{P}^*)$	covariance matrix
$D_j$	contribution of second-order group of type- $j$
$E_k$	contribution of third-order group of type- $k$
$f(X)$	function of property $X$
GC	group-contribution
MG	Marrero and Gani
$M_j$	occurrence of second-order group of type- $j$
N	number of experimental data-points used in the regression
$N_i$	occurrence of first-order group of type- $i$
$O_k$	occurrence of third-order group of type- $k$
$\mathbf{P}$	model parameters

$P^*$	estimated values of model parameters
$R^2$	coefficient of determination
$S(P)$	cost function
SD	standard deviation
$SSE$	minimum sum of squared errors
$t(v, \alpha_t/2)$	t-distribution value corresponding to the $\alpha_t/2$ percentile
$v$	degrees of freedom
$X^{exp}$	experimental property value
$X^{pred}$	predicted property value

**Highlights**

- Prediction of normal melting point, water solubility and octanol-water partition coefficient of amino acids
- Group contribution approach to develop the property models
- Provides uncertainty estimates for the modeled properties
- Provides modelling details together with model parameters